A comparative Study of Locality Preserving Projection and Principle Component Analysis on Classification Performance using Logistic Regression

Azza kamal Ahmed Abdelmajed

B.Sc. (Honours) in Statistics/Computer, University of Gezira, (2009)

A Dissertation

Submitted to the University of Gezira in Partial Fulfillment of the Requirements for the Award of the Degree of Master of Science

in

Computer Science

Department of Computer Science

Faculty of Mathematical and Computer Sciences

November, 2015
A comparative Study of Locality Preserving Projection and Principle Component Analysis on Classification Performance using Logistic Regression

Azza kamal Ahmed Abdelmajed

Supervision Committee:

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Signature</th>
</tr>
</thead>
</table>
| Dr. Murtada Khalfallah Elbashir | Main Supervisor | ...........
| Dr. Abdallah Bashir Musa  | Co-supervisor  | ...........

Date: November, 2015
A comparative Study of Locality Preserving Projection and Principle Component Analysis on Classification Performance using Logistic Regression

Azza Kamal Ahmed Abdemajed

Examination Committee:

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Murtada Khalfallah Elbashir</td>
<td>ChairPerson</td>
<td>..........</td>
</tr>
<tr>
<td>Dr. Abdalla Akod Osman</td>
<td>External Examiner</td>
<td>..........</td>
</tr>
<tr>
<td>Dr. Sally Dafa Allah Awad Elkarim</td>
<td>Internal Examiner</td>
<td>..........</td>
</tr>
</tbody>
</table>

Date of Examination: 18 / 11 / 2015
DECLARATION

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. The work was done under the guidance of doctor Murtada Khalafalla Elbashir, dean of faculty of mathematical and computer science and doctor Abdalla Bashir Musa, dean of informatics administration.

Azza Kamal Ahmed Abdelmajed
DEDICATION

I would like to dedicate this dissertation to my mother, Batoul Ahmed Altayeb. There is no doubt in my mind that without her continued love, support and counsel I could not have completed this work.

I dedicate this work and give special thanks to my family and many friends. A special feeling of gratitude to my sisters Asgid, Abeer and Hadeel, have never left my side and are very special.

I also dedicate this dissertation to my brother Abass and my family who have supported me throughout the process.
ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Dr. Murtada Khalfallah Elbashir Dean of Faculty of Mathematical and Computer Sciences -Gezira University- Sudan, for his guidance, encouragement, suggestions and support during the progress of the research and realization of the research.

I would also like to give my sincere gratitude to Dr. Abdallah Bashir Musa, Dean of Informatics Administration -University of Gezira- Sudan, for his support and help.
A comparative Study of Locality Preserving Projection and Principle Component Analysis on Classification Performance using Logistic Regression

Azza Kamal Ahmed Abdelmajed

Abstract

Data mining is the extraction and retrieval of useful data and also involves the retrieval and analysis of data that is stored in a data warehouse. Some of the major techniques of data mining are classification, association and clustering. Data mining is an upcoming research area to solve various problems and classification is one of them. There are many classification techniques such as neural network, decision tree, support vector machine and logistic regression. The problem of dimensionality is pertinent to many learning algorithms, and it denotes the drastic raise of computational complexity and the classification error in high dimensions, thus the dimensions need to be reduced using dimensionality reduction algorithms. The problem of this research is in many real-world classification problems, the local structure is more important than the global structure, and in the dimensionality reduction algorithms such as principal component analysis (PCA) it preserves the global structure of the dataset and ignores the local structure of the dataset. Therefore this research will resolve this problem using Locality Preserving Projections (LPP) algorithm that is preserving the local structure of the datasets. LPP is a linear projective maps that arise by solving variational problem that optimally preserves the neighborhood structure of the data set. LPP should be seen as an alternative to principal component analysis (PCA) and also not found study using LPP algorithm with logistic regression model. The aims of this research are to compare between PCA and LPP in terms of accuracy, to develop appropriate representations of complex data by reducing the dimensions of the data and to explain the importance of using LPP with logistic regression. The methodology of this research compared the proposed LPP approach with PCA method on five different data sets using dimensionality reduction toolbox (drtoolbox) in matlab software and evaluation the model using cross validation method and then calculated the performance measures (accuracy, sensitivity, Specificity, precision, f-score and roc curve) of both. The experimental results of this research find that the proposed LPP approach provides a better representation and high accuracy than the PCA approach. This research recommends by using more than mathematical model such as neural network, decision tree and increase the data sets more than five to obtain the best results.
دراسة مقارنة بين طريقة الاسقاطات المحافظة على المحلية وطريقة تحليل المكون الرئيسي على اداء التصنيف باستخدام الانحدار اللوجستي

عزه كمال احمد عبد الماجد

ملخص الدراسة

التنقيب على البيانات هو عبارة عن استخلاص واسترجاع البيانات المفيدة ويتضمن أيضا استرجاع وتحليل البيانات التي تم تخزينها في مخزن البيانات. بعض من التقنيات الرئيسية للتنقيب عن البيانات هي التصنيف الإرباط والمجموعات. ويعتبر التنقيب عن البيانات مجال جديد لحل عدد من المشاكل والتقييم هو أحد هذه المشاكل. هناك تقنيات تصنيف مختلفة مثل الشبكات العصبية، شجرة القرار، المتجددة الانحدار ولوجستي. تعتبر مشكلة الانقراض ذات صلة وثيقة بعدم خوارزميات التعلم وتفرز البيانات في التطبيق الحسابي، و.quick التصنيف في الابعاد العالية، لذلك نحتاج أن نقل الابعاد باستخدام خوارزميات تخفيف الابعاد. مشكلة هذا البحث تتمثل في أنه في العديد من مشاكل تصنيف العالم الحقيقي يعتبر التركيب المحلي للبيانات أكثر أهمية من التركيب العالمي. وفي خوارزميات تخفيف الابعاد مثل خوارزمية تحليل المكون الرئيسي التي تحافظ على التركيب العالمي لمجموعه البيانات وتجاهل التركيب المحلي لها، لذلك هذا البحث حل هذه المشكلة باستخدام خوارزمية الأسقاطات المحافظة على المحلية التي تحافظ على التركيب المحلي لمجموعه البيانات، وهي عبارة عن الخرائط الإسفنجية الخطية والتي ظهرت بحل مشكلة التجانس والتي تحافظ بصوره مثالية على التركيب المجاور لمجموعه البيانات. الأسقاطات المحافظة على المحلية يجب أن ينظر لها على أنها بديل لخوارزمية تحليل المكون الرئيسي، وأيضًا لا توجد دراسة تستخدم هذه الخوارزمية مع نموذج الانحدار اللوجستي. اهداف هذا البحث تتمثل في المقارنة بين خوارزمية الأسقاطات المحافظة على المحلية وخوارزمية تحليل المكون الرئيسي من حيث الدقة، تطوير مثاليات مناسبة من البيانات المعقدة عن طريق الحد من أبعاد البيانات ووضع أهمية استخدام الخوارزمية المحافظة على المحلية مع الانحدار اللوجستي. منهجية هذا البحث قارنت بين الخوارزميتين اعلاه على خمس مجموعات بيانات مختلفة باستخدام صندوق ألوان تخفيض الابعاد في برنامج الماتلاب، وتم تقييم النموذج باستخدام درجة التحقق من الصحة، ثم حساب معيار الاداء (المعطيات، محدودة، معيار F). منحنى روك (لكل الخوارزميتين توصلت النتائج التجريبية لهذا البحث أن خوارزميات الأسقاطات المحافظة على المحلية تقدم إعادة تمثيل أفضل ودقة عالية للبيانات أكثر من خوارزمية تحليل المكون الرئيسي. يوصي هذا البحث باستخدام أكثر من نموذج رياضي مثل الشبكات العصبية وشجرة القرار وزيادة مجموعات البيانات إلى أكثر من خاصة وذلك للحصول على نتائج أفضل للخوارزمية.

منهجية الدراسة

هذا البحث قارن بين خوارزميتين اعلاه على خمسة مجموعات بيانات مختلفة باستخدام مجموع الألوان تخفيض الابعاد في برنامج الماتلاب، وتم تقييم النموذج باستخدام درجة التحقق من الصحة، ثم حساب معيار الاداء. (المعطيات، محدودة، معيار F). منحنى روك (لكل الخوارزميتين) توصلت النتائج التجريبية. لهذا البحث أن خوارزميات الأسقاطات المحافظة على المحلية تقدم إعادة تمثيل أفضل ودقة عالية للبيانات أكثر من خوارزمية تحليل المكون الرئيسي. يوصي هذا البحث باستخدام أكثر من نموذج رياضي مثل الشبكات العصبية، وشجرة القرار وزيادة مجموعات البيانات إلى أكثر من خمسة وذلك للحصول على نتائج أفضل للخوارزمية.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Subject</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declaration</td>
<td>IV</td>
</tr>
<tr>
<td>Dedication</td>
<td>V</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>VI</td>
</tr>
<tr>
<td>Abstract</td>
<td>VII</td>
</tr>
<tr>
<td>Abstract (in Arabic)</td>
<td>VII</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>VIII</td>
</tr>
<tr>
<td>List of Tables</td>
<td>XI</td>
</tr>
<tr>
<td>List of Figures</td>
<td>XII</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>XIII</td>
</tr>
</tbody>
</table>

## CHAPTER 1: INTRODUCTION

1.1 Introduction                        | 1   |
1.2 Problem Definition                 | 2   |
1.3 Research Objectives                | 3   |
1.4 Research Methodology               | 3   |
1.5 Structure of the research          | 3   |

## CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

2.1 Background                         | 4   |
2.1.1 Data mining                      | 4   |
2.1.2 Machine learning ................................................................. 5
2.1.3 Classification ....................................................................... 7
2.1.4 Logistic Regressions............................................................... 8
2.1.4.1 Binary Logistic Regressions .............................................. 8
2.1.4.2 Multiple Logistic Regressions .......................................... 9
2.1.4.3 Maximum Likelihood ..................................................... 9
2.1.5 Dimensionality reductions .................................................. 10
2.1.6 The concept of locality preserving projections (LPP) ............ 10
2.1.6.1 The linear dimensionality reduction problem .................. 11
2.1.6.2 The algorithm .............................................................. 11
2.1.7 The concept of principle component analysis (PCA) ............ 12
2.1.8 Previous studies .................................................................. 14
2.1.8.1 Literature Review Summary ........................................... 16

CHAPTER 3: METHODOLOGY

3.1 Introduction ........................................................................... 18
3.2 Experimental Setup ............................................................... 19
3.2.1 dimensionality reduction toolbox (drtoolbox) ....................... 20
3.2.2 Cross validation ............................................................... 20
3.2.3 Performance Measure ....................................................... 21
3.3 Interfaces ............................................................................. 23

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction ........................................................................... 26
4.2 Results ................................................................................ 26
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table name</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1: Confusion matrix</td>
<td>21</td>
</tr>
<tr>
<td>Table 4.1: The results of the performance measures for LR with all the variables</td>
<td>26</td>
</tr>
<tr>
<td>Table 4.2: The results of the performance measures for LR –PCA</td>
<td>27</td>
</tr>
<tr>
<td>Table 4.3: The results of the performance measures for LR –LPP</td>
<td>28</td>
</tr>
<tr>
<td>Table 4.4: Area under the curve</td>
<td>33</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure name</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1: Data mining stages</td>
<td>5</td>
</tr>
<tr>
<td>Figure 3.1: Methodology</td>
<td>18</td>
</tr>
<tr>
<td>Figure 3.2: Cross validation steps</td>
<td>20</td>
</tr>
<tr>
<td>Figure 3.3: Import data</td>
<td>24</td>
</tr>
<tr>
<td>Figure 3.4: data without Y</td>
<td>25</td>
</tr>
<tr>
<td>Figure 3.5: PCA or LPP</td>
<td>26</td>
</tr>
<tr>
<td>Figure 3.6: Cross validation</td>
<td>26</td>
</tr>
<tr>
<td>Figure 4.1: Roc curve for climate model simulation crashes data</td>
<td>29</td>
</tr>
<tr>
<td>Figure 4.2: Roc curve for heart data set</td>
<td>30</td>
</tr>
<tr>
<td>Figure 4.3: Roc curve for Spam base data set</td>
<td>31</td>
</tr>
<tr>
<td>Figure 4.4: Roc curve for phishing websites data set</td>
<td>32</td>
</tr>
<tr>
<td>Figure 4.5: Roc curve for musk (version 1) data set</td>
<td>33</td>
</tr>
</tbody>
</table>
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Area under curve</td>
</tr>
<tr>
<td>DM</td>
<td>Data Mining</td>
</tr>
<tr>
<td>DR</td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td>DRtoolbox</td>
<td>Dimensionality reduction toolbox</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>LPP</td>
<td>Locality Preserving Projections</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>MLR</td>
<td>Multiple Logistic Regression</td>
</tr>
<tr>
<td>PPV</td>
<td>Positive predictive value</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>TN</td>
<td>False Negative</td>
</tr>
</tbody>
</table>
Introduction

1.1 Introduction

Data mining is the extraction and retrieval of useful data and also involves the retrieval and analysis of data that is stored in a data warehouse. Some of the major techniques of data mining are classification, association and clustering. Data mining is an upcoming research area to solve various problems and classification is one of the main problems in the field of data mining [1]. Classification predicts categorical class labels and it classifies the data based on the training set and the values in classifying the attributes and uses it in classifying the new data. Data classification is a two-step process consisting of model construction and model usage. Model construction is used for describing predetermined classes. Model usage is used for classifying future or unknown objects [2]. The development of data mining applications such as classification has shown the need for machine learning algorithms to be applied specially in large scale data [3].

Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data, there is two type of machine learning: supervised learning and unsupervised learning, the first one is use training data to infer model and apply model to test data and for the other one there is no training data and model inference and application both rely on test data exclusively. Modern machine learning techniques are progressively being used by biologists to obtain proper results from the databases.

There are a variety of classification techniques such as neural network, decision tree, support vector machine and logistic regression.

Logistic regression (LR) is a well-known statistical classification method that has been used widely in a variety of applications including document classification, bioinformatics and analyzing a data set in which there are one or more independent variables that determine an outcome[4]. The advantage of using LR is that deal with dependent variables that are categorical and extension to the multi class case. Logistic regression not just classes the data but also predicts probabilities and then estimate the parameters of the model using maximum likelihood estimator.

Before using the data set in the classification it needs some preprocesses such as data cleaning, data transformation and data reduction, the last one is very important because usually represent the dataset in an (n*m) dimensional space, these(n*m)
dimensional spaces are too large, however need to reduce the size of the data set before applying a learning algorithm. A common way to attempt to resolve this problem is to use dimensionality reduction techniques [5].

Because principle component analysis (PCA) it preserves the global structure of the data set and ignores the local structure of the data set. Therefore this research proposes linear dimensionality reduction algorithm, called Locality Preserving Projections (LPP). LPP is linear projective maps that arise by solving variational problem that optimally preserves the neighborhood structure of the data set, LPP should be seen as an alternative to principal component analysis (PCA) [6].

1.2 Research problem:

In the dimensionality reduction algorithms such as principle component analysis (PCA) it preserves the global structure of the data set and ignores the local structure of the data set and the local structure is more important than the global structure for many reason:

• It is important to maintain the intrinsic information of high-dimensional data when they are transformed to a low dimensional space for analysis.

• A single characterization, either global or local, may be insufficient to represent the underlying structures of real world data.

• The local geometric structure of data can be seen as a data dependent regularization of the transformation matrix, which helps to avoid over fitting, especially when training samples are scarce.

1.3 Objectives:

1/ To develop appropriate representations of complex data by reducing the dimensions.

2/ To explain the importance of using locality preserving projection algorithm with Logistic regression.

3/ To compare between (LPP) and (PCA) in terms of performance measures.
1.4 Methodology:

In this research five data sets were be used and then apply the algorithms principle component analysis (PCA) and Locality Preserving Projections (LPP) to reduce the dimensions of the data using dimensionality reduction toolbox (drtoolbox) in matlab software. After the input space is reduced to a lower dimension by applying one of the two methods PCA and LPP, cross-validation method will be applied to this new reduced features space using 10 fold to evaluation the model and then apply logistic regression to classifier the reduced data. All the performance measures: accuracy, sensitivity, specificity, f-score, precision and roc curve will be computed. The ROC analysis is plotted after each cross validation for the two methods using spss software to compute the area under the curve.

1.5 Structure of the research:

This research is consists of five chapters, it starts with introduction, problem identification and justification followed by objectives, and ends with Structure of the research, all presented in chapter one. Chapter two related concepts and reviews the theoretical background and information about the logistic regression and locality preserving projection algorithm which will be used in this research. Chapter three consists of methodology. Chapter four application screens are explained and discuss the results of the reality of the analytical study, and then the fifth chapter conclusion and recommendations.
Background and Literature Review

2.1 Background

This chapter talks about the background theory for basic concept of this research like: data mining, machine learning, classification, logistic regression, dimensionality reduction, principle component analysis, locality preserving projection and also talk about the previous studies.

2.1.1 Data mining

Data Mining is commonly defined as the computer assisted search for interesting patterns and relations in large databases[7], it is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction and predictive data mining is the most common type of data mining.

The process of data mining consists of three stages shown in figure 2.1:

(1) Exploration.

(2) Model building and validation.

(3) Deployment.

**Stage 1: Exploration**: This stage involves preparation and collection of data, it also involves data cleaning, data transformation and data reduction. Based on size of data, different tools to analyze the data may be required. This stage helps to determine different variables of the data to determine their behavior.

**Stage 2: Model building and validation**: This stage involves choosing the best model based on their predictive performance. The model is then applied on the different data sets and compared for best performance.

**Stage 3: Deployment**: Based on model selected in previous stage, it is applied to the data sets. This is to generate predictions or estimates of the expected outcome.

The concept of data mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that
can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business data mining but data mining is still based on the conceptual principles of statistics including the traditional exploratory data analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques[8].

![Data mining stages diagram](image)

**Figure 2.1 Data mining stages**

The development of data mining applications such as classification has shown the need for machine learning algorithms to be applied specially in large scale data.

**2.1.2 Machine Learning (ML)**

ML is generally taken to involve automatic computing procedures based on logical or computational learning theory, and similar terms are often used in the context of data mining, to denote the application of generic model fitting or classification algorithms for predictive data mining. Unlike traditional statistical data analysis, which is usually concerned with the estimation of population parameters by statistical inference, the emphasis in data mining (and machine learning) is usually on the accuracy of prediction, regardless of whether or not the "models" or techniques that are used to generate the prediction is interpretable or open to simple explanation.
Good examples of this type of technique often applied to predictive data mining are neural networks or meta learning techniques. These methods usually involve the fitting of very complex "generic" models that are not related to any reasoning or theoretical understanding of underlying causal processes, instead, these techniques can be shown to generate accurate predictions or classification in cross validation samples. Machine learning can be considered a subfield of computer science and statistics. It has strong ties to artificial intelligence and optimization, which deliver methods, theory and application domains to the field.

Machine learning tasks are typically classified into three broad categories and those are:

- **Supervised learning:**
  
  Supervised learning entails learning a mapping between a set of input variables $X$ and output variable $Y$ and applying this mapping to predict the outputs for unseen data. Supervised learning is the most important methodology in machine learning and it also has a central importance in the processing of multimedia data [10].

- **Unsupervised learning:**
  
  Studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns and it can be a goal in itself (discovering hidden patterns in data) or a means towards an end [11].

- **Reinforcement learning:**
  
  Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them.[12].

Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system.

- In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one (or multi-label
classification) or more of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".

- In regression, also a supervised problem, the outputs are continuous rather than discrete.
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
- Density estimation finds the distribution of inputs in some space.
- Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. [13]

### 2.1.3 Classification:

Classification is one of the data mining techniques that is mainly used to analyze a given data set and takes each instance of it and assigns this instance to a particular class with the aim of achieving least classification error. It is used to extract models that correctly define important data classes within the given dataset. It is a two step process. In first step the model is created by applying classification algorithm on training data set. Then in second step, the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy [14]. It is a data mining task of predicting the value of a categorical variable by building a model based on one or more numerical and/or categorical variables (predictors). Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how “good” the algorithm is.

There are a variety of classification techniques such as neural network, decision tree, support vector machine and logistic regression. In this research, has been present a logistic regression for the classification.
2.1.4 Logistic regression:

Logistic regression (LR) is a well known statistical classification method that has been used widely in a variety of applications including document classification, bioinformatics and analyzing a data set in which there are one or more independent variables that determine an outcome. The advantage of using LR is that deal with dependent variables that are Categorical and extension to the multi class case.

Let Y be a random variable such that

\[
Y = \begin{cases} 
1 & \text{if the condition is present} \\
0 & \text{otherwise} 
\end{cases}
\]

2.1.4.1 Binary Logistic Regression:

The Binary Logistic Regression model is given by:

\[
P(X) = \frac{\exp(B_0 + B_1 X_1)}{1 + \exp(B_0 + B_1 X_1)} \tag{2.1}
\]

Where B0, B1 are a parameters of logistic regression model.

2.1.4.2 Multi Logistic Regression:

now consider the problem of predicting a binary response using multiple predictors. By analogy with the extension from simple to multiple linear regression, we can generalize (2.1) as follows:

\[
\text{Log} \left( \frac{P(X)}{1-P(X)} \right) = B_0 + B_1 X_1 + B_2 X_2, \ldots + B_p X_p \tag{2.2}
\]

Where \( X = (X_1, X_2, \ldots, X_p) \) are predictors. Equation 2.2 can be rewritten as be covariates of interest. Define:
If $\beta^*$ be the maximum likelihood estimation of $\beta=(B_0,B_1,\ldots,B_p)$ then the probability that a new observation $X^*$=$(x_1^*,x_2^*,\ldots,x_p^*)$ define as [15]:

$$P^*(x^*) = \frac{\text{Exp}(B_0^*+B_1^*x_1^*+\cdots+B_p^*x_p^*)}{1+\text{Exp}(B_0^*+B_1^*x_1^*+\cdots+B_p^*x_p^*)} \quad (2.4)$$

### 2.1.4.3 Maximum Likelihood

**Maximum Likelihood.** Instead of finding the best fitting line by minimizing the squared residuals, as we did with OLS regression, we use a different approach with logistic—Maximum Likelihood (ML). ML is a way of finding the smallest possible deviance between the observed and predicted values (kind of like finding the best fitting line) using calculus (derivatives specifically). With ML, the computer uses different "iterations" in which it tries different solutions until it gets the smallest possible deviance or best fit. Once it has found the best solution, it provides a final value for the deviance, which is usually referred to as "negative two log likelihood" [16].

### 2.1.5 Dimensionality Reductions:

In many real-world applications, numerous features are used in an attempt to ensure accurate classification. If all those features are used to build up classifiers, then they operate in high dimensions, and the learning process becomes computationally and analytically complicated, resulting often in the drastic rise of classification error. Hence, there is a need to reduce the dimensionality of the feature space before classification. According to the adopted strategy dimensionality reduction techniques are divided into feature selection and feature transformation. The key difference between feature selection and feature transformation is that during the first process a subset of original features only is selected while the second approach is based on the
generation of completely new features[17]. Feature extraction is a dimensionality reduction technique that extracts a subset of new features from the original set of features by means of some functional mapping keeping as much information in the data as possible [18].

Dimensionality reduction (DR) is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation has a dimensionality that corresponds to the intrinsic dimensionality of the data. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data. Dimensionality reduction is important in many domains, since it facilitates classification, visualization, and compression of high-dimensional data, by mitigating the curse of dimensionality and other undesired properties of high-dimensional spaces.

2.1.6 The Concept of Locality Preserving Projection (LPP):

Many problems in information processing involve some form of dimensionality reduction. In this research locality preserving projections (LPP) algorithm is introduced. These are linear projective maps that arise by solving a variational problem that optimally preserves the neighborhood structure of the data set. LPP should be seen as an alternative to principal component analysis (PCA) – a classical linear technique that projects the data along the directions of maximal variance. When the high dimensional data lies on a low dimensional manifold embedded in the ambient space, the Locality Preserving Projections are obtained by finding the optimal linear approximations to the Eigen functions of the Laplace Beltrami operator on the manifold.

Statistical analysis of the LPP algorithm, different from Principal Component Analysis (PCA) which obtains a subspace spanned by the largest eigenvectors of the global covariance matrix, show that LPP obtains a subspace spanned by the smallest eigenvectors of the local covariance matrix.

It is worthwhile to highlight several aspects of the proposed approach here:

1. While PCA aims to preserve the global structure of the image space, and LDA aims to preserve the discriminating information; LPP aims to preserve the local structure of the image space. In many real world classification problems, the local manifold
structure is more important than the global Euclidean structure, especially when nearest-neighbor like classifiers are used for classification.

2. An efficient subspace learning algorithm for face recognition should be able to detect the nonlinear manifold structure of the face space.

3. LPP is obtained by finding the optimal linear approximations to the eigen functions of the Laplace Beltrami operator on the manifold.

LPP is linear. Moreover, LPP is defined everywhere, LPP may be simply applied to any new data point to locate it in the reduced representation space.

2.1.6.1 The Linear Dimensionality Reduction Problem:

The generic problem of linear dimensionality reduction is the following. Given a set \(x_1, x_2, ... x_m\) in \(\mathbb{R}^l\), find a transformation matrix \(A\) that maps these \(m\) points to a set of points \(y_1, y_2, ..., y_m\) in \(\mathbb{R}^n\) (\(l < n\)), such that \(y_i\) "represents" \(x_i\), where \(y_i = A^T x_i\).

This method is of particular applicability in the special case where \(x_1, x_2, ..., x_m, M\) and \(M\) is a nonlinear manifold embedded in \(\mathbb{R}^n\).

2.1.6.2 The algorithm

Locality Preserving Projection (LPP) is a linear approximation of the nonlinear Laplacian Eigenmap. The algorithmic procedure is formally stated below:

1. **Constructing the adjacency graph:**
   Let \(G\) denote a graph with \(m\) nodes. put an edge between nodes \(i\) and \(j\) if \(x_i\) and \(x_j\) are "close".

   **There are two variations:**
   (a) \(\varepsilon\)- neighborhoods. [parameter \(\varepsilon \in \mathbb{R}\)] Nodes \(i\) and \(j\) are connected by an edge if \(\|x_i - x_j\| < \varepsilon\) where the norm is the usual Euclidean norm in \(\mathbb{R}^n\).
   (b) \(k\) nearest neighbors. [parameter \(k \in \mathbb{N}\)] Nodes \(i\) and \(j\) are connected by an edge if \(i\) is among \(k\) nearest neighbors of \(j\) or \(j\) is among \(k\) nearest neighbors of \(i\).

2. **Choosing the Weights:**
   Here, as well, I have two variations for weighting the edges. \(W\) is a sparse symmetric \(m \times m\) matrix with \(W_{ij}\) having the weight of the edge joining vertices \(i\) and \(j\), and 0 if there is no such edge.
   (a) Heat kernel. [parameter \(t \in \mathbb{R}\)]. If nodes \(i\) and \(j\) are connected, put
\[ W_{ij} = e^{-||x_i-x_j||^2/2} \quad \ldots \ldots \text{(2.5)} \]

(b) Simple-minded. [No parameter]. \( W_{ij} = 1 \) if and only if vertices \( i \) and \( j \) are connected by an edge.

3. Eigenmaps:
Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:

\[ X L X^T a = \lambda X D X^T a \quad \ldots \ldots \text{(2.6)} \]

where \( D \) is a diagonal matrix whose entries are column (or row, since \( W \) is symmetric) sums of \( W \), \( D_{ii} = \sum_j W_{ji} \). \( L = D - W \) is the Laplacian matrix. The \( i^{th} \) column of matrix \( X \) is \( x_i \).

Let the column vectors \( a_0, a_1, \ldots, a_{i-1} \) be the solutions of equation, ordered according to their eigenvalues, \( \lambda_0 < \lambda_1 \ldots < \lambda_{i-1} \). Thus, the embedding is as follows:

\[ \rightarrow x_i \quad y_i = A^T x_i \quad , \quad A = (a_0, 1, \ldots, a_{i-1}) \]

where \( y_i \) is a \( I \)-dimensional vector, and \( A \) is a \( n \times I \) matrix [6].

2.1.7 The concept of principle component analysis (PCA)

Principal components analysis (PCA) is used for two objectives:
1. Reducing the number of variables comprising a dataset while retaining the variability in the data.
2. Identifying hidden patterns in the data, and classifying them according to how much of the information, stored in the data, they account for.

Properties of the PCA:

(1) it maximizes the variance of the extracted features.
(2) the extracted features are uncorrelated.
(3) it finds the best linear approximation.
(4) it maximizes the information contained in the extracted features.
2.1.7.1 The Computation of the PCA

1) Calculate the covariance matrix from the input data.

\[ \text{COV}(X,Y) = \sum \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{N} \] \hspace{1cm} \text{(2.7)}

where

- \( N \) is the number of scores in each set of data
- \( \bar{X} \) is the mean of the \( N \) scores in the first data set
- \( X_i \) is the \( i \)th raw score in the first set of scores
- \( \bar{Y} \) is the mean of the \( N \) scores in the second data set
- \( Y_i \) is the \( i \)th raw score in the second set of scores
- \( \text{Cov}(X, Y) \) is the covariance of corresponding scores in the two sets of data

2) Compute the eigenvalues and eigenvectors of \( S \) and sort them in a descending order with respect to the eigenvalues. Let \( A \) be an \( n \times n \) matrix. The eigenvalues of \( A \) are defined as the roots of:

\[ \text{Determinant} \ (A - \lambda I) = | (A - \lambda I) | = 0 \] \hspace{1cm} \text{(2.8)}

Where \( I \) is the \( n \times n \) identity matrix. Let \( \lambda \) be an eigenvalue of \( A \). Then there exists a vector \( x \) such that:

\[ Ax = \lambda x \] \hspace{1cm} \text{(2.9)}

The vector \( x \) is called an eigenvector of \( A \) associated with the eigenvalue. Notice that there is no unique solution for \( x \) in the above equation. It is a direction vector only and can be scaled to any size.

3) Form the actual transition matrix by taking the predefined number of components (eigenvectors).

4) Chosen the components (eigenvectors) that the eigenvector with the highest eigenvalue is the principle component of the data set
5) Finally, multiply the original feature space with the obtained transition matrix, which yields a lower-dimensional representation.

This process is equivalent to finding the axis system in which the covariance matrix is diagonal.

Variation, the one with the Second largest eigenvalue is the (orthogonal) direction with the next highest variation and so on.

### 2.1.8 Previous studies:

**In 2003**, Xiaofei He*, Shuicheng Yan#, Yuxiao Hu, and Hong-Jiang Zhang, propose a new approach to mapping face images into a subspace obtained by Locality Preserving Projections (LPP) for face analysis. We call this Laplacianface approach. The results of this paper found the Laplacianface approach provides a better representation and achieves lower error rates in face recognition[30].

**In 2004**, Yun Tang and Richard Rose, presents a new approach to feature analysis in automatic speech recognition (ASR) based on locality preserving projections (LPP). The results of this Paper obtained on the Resource Management (RM) data set showed that when LPP based dimensionality reduction was applied in the context of mel frequency cepstrum coefficient (MFCC) based feature analysis, a significant reduction of word error rate (WER) was obtained with respect to standard MFCC features and a High dimensional vectors obtained by concatenating consecutive static feature vectors are projected to a low dimensional subspace[31].

**In 2005**, Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang, Fellow, propose an appearance-based face recognition method called the Laplacianface approach. The result of this Paper found the proposed Laplacianface approach provides a better representation and achieves lower error rates in face recognition[5].

**In 2007**, Zhonglong Zhenga,, Fan Yanga, Wenan Tana, Jiong Jiaa, Jie Yangb, introduces a novel Gabor-based supervised locality preserving projection (GSLPP) method for face recognition. The result of this paper found the AR database and CMU PIE database show superior of the novel GSLPP method[32].

**In 2010**, Shermina.J focuses on a systematic analysis of locality-preserving projections and the application of LPP in combination with an existing technique. The
result of this Paper show the significant improvements in the face recognition performance in comparison with some previous methods[33].

In 2011, Jinchao Yang, Xiang Zhang, Li Lu, Jianping Zhang, Yonghong Yan, introduce locality preserving projection (LPP) to language recognition under the support vector machine (SVM) frame work. The result of this paper found the new language features of Laplacian super vector that preserve local structure and nonlinear manifolds also contain discriminative language dependent information[34].

In 2012, Wei Li, Student Member, IEEE, Saurabh Prasad, Member, IEEE, James E. Fowler, Senior Member, IEEE, and Lori Mann Bruce, Senior Member, IEEE, presented a locality-preserving discriminant analysis for hyperspectral dimensionality reduction and propose a classification paradigm that is designed to exploit the rich statistical structure of the data. The result of this paper found the LFDA-GMM/SVM approach to be even more effective at capturing subtle statistical differences for classification in such complex environments[35].

In 2014, Le Shu, Tianyang Ma, Longin Jan Latecki, proposes a novel locality preserving projection method for domain adaptation task. The result of this Paper found that the new feature representation which can effectively preserve the local structure[36].

2.1.8.1 Literature Review Summary:

Table 2.1 shows the literature review summary
Table 2.1 The Literature Review Summary

<table>
<thead>
<tr>
<th>STUDY TITLE</th>
<th>TECHNOLOGY</th>
<th>RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning a Locality Preserving Subspace for Visual Recognition</td>
<td>Locality Preserving Projections (LPP) for face analysis</td>
<td>Laplacianface approach provides a better representation and achieves lower error rates in face recognition</td>
</tr>
<tr>
<td>A Study of using Locality Preserving Projections for Feature Extraction in Speech Recognition</td>
<td>new approach to feature analysis in automatic speech recognition (ASR) based on locality preserving projections (LPP)</td>
<td>a significant reduction of word error rate (WER) was obtained with respect to standard MFCC features and a High dimensional vectors obtained by concatenating consecutive static feature vectors are projected to a low dimensional subspace</td>
</tr>
<tr>
<td>Face Recognition Using Laplacianfaces</td>
<td>an appearance-based face recognition method called the Laplacianface approach</td>
<td>Laplacianface approach provides a better representation and achieves lower error rates in face recognition</td>
</tr>
<tr>
<td>Gabor feature based face recognition using supervised locality preserving projection</td>
<td>A novel Gabor-based supervised locality preserving projection (GSLPP) method for face recognition</td>
<td>the AR database and CMU PIE database show superior of the novel GSLPP method</td>
</tr>
<tr>
<td>Application of Locality Preserving Projections in Face Recognition</td>
<td>systematic analysis of locality-preserving projections and the application of LPP in combination with an existing technique</td>
<td>the significant improvements in the face recognition performance in comparison with some previous methods</td>
</tr>
<tr>
<td>Language Recognition With Locality Preserving Projection</td>
<td>locality preserving projection (LPP) to language recognition under the support vector machine (SVM) frame work</td>
<td>the new language features of Laplacian super vector that preserve local structure and nonlinear manifolds also contain discriminative language dependent information</td>
</tr>
<tr>
<td>Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis</td>
<td>locality-preserving discriminant analysis for hyperspectral dimensionality reduction</td>
<td>the LFDA-GMM/SVM approach to be even more effective at capturing subtle statistical differences for classification in such complex environments</td>
</tr>
<tr>
<td>Locality Preserving Projection for Domain Adaptation with Multi-Objective Learning</td>
<td>A novel locality preserving projection method for domain adaptation task.</td>
<td>the new feature representation which can effectively preserve the local structure[</td>
</tr>
</tbody>
</table>

All previous studies above used locality preserving projection with different methods, but there is no study used LPP with logistic regression. This research used LPP with logistic regression.
METHODOLOGY

3.1 Introduction:

In this research locality preserving projection (LPP) algorithm for dimensionality reduction is used on the data sets and compare it with principle component analysis (PCA) algorithm using dimensionality reduction toolbox in matlab as a tool and apply logistic regression technique for classification and then calculate the performance measures of both.

![Figure 3.1 Methodology](image)

The figure 3.1 explain the methodology of this research.
3.2 Experimental Setup:

In order to evaluate a prediction method it is necessary to have different data sets for training and testing, however five datasets will be used and apply the algorithms principle component analysis (PCA) and Locality Preserving Projections (LPP) to reduce the dimensions using dimensionality reduction toolbox (drtoolbox) in matlab software. After the input space is reduced to a lower dimension by applying one of the two methods PCA and LPP, cross-validation method will be applied to this new reduced features space using 10 fold to evaluation the model and then apply logistic regression to classifier the reduced data. All the performance measures: accuracy, sensitivity, specificity, f-score, precision and roc curve will be computed. The ROC analysis is plotted after each cross validation for the two methods using spss software to compute the area under the curve.

3.2.1 Dimensionality Reduction Toolbox (drtoolbox):

Is the matlab toolbox for dimensionality reduction there are free to use, change, or redistribute this code in any way for non-commercial purposes. This Matlab toolbox implements 34 techniques for dimensionality reduction and metric learning [24]. Furthermore, the toolbox contains 6 techniques for intrinsic dimensionality estimation.

3.2.2 Cross validation:

Cross validation is a model evaluation method that is better than residuals. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on "new" data. This is the basic idea for a whole class of model evaluation methods called cross validation [25].

In this research split the data set into 10 folds as show in figure 3.2.
In the first step take fold 1 as a test and the rest of the data as a train and then repeat this process for the next fold, until all folds have been used to either train or test the model.

Figure 3.2: Cross validation steps
3.2.3 Performance Measure:

The measures that are used in this research depend on matrix called the confusion matrix are as follows in table 3.2:

<table>
<thead>
<tr>
<th></th>
<th>Predicted positive</th>
<th>Predicted negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual positive</td>
<td>TP</td>
<td>FN</td>
<td>AP</td>
</tr>
<tr>
<td>Actual negative</td>
<td>FP</td>
<td>TN</td>
<td>AN</td>
</tr>
<tr>
<td>Total</td>
<td>PP</td>
<td>PN</td>
<td>N</td>
</tr>
</tbody>
</table>

Where:

TP: true positives (predicted positive, actual positive)

TN: true negatives (predicted negative, actual negative)

FP: false positives (predicted positive, actual negative)

FN: false negatives (predicted negative, actual positive)[26].

- **Accuracy:**

Accuracy is the proportion of true results (both true positives and true negatives) in the population [27].

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.........(3.1)}
\]

- **Sensitivity or Recall:**

Proportion of actual positives which are predicted positive[27].

\[
\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{.........(3.2)}
\]
• **Specificity:**

proportion of actual negative which are predicted negative[27].

\[
\text{Specificity} = \frac{TN}{TN + FP} \quad \ldots \ldots (3.3)
\]

• **Positive predictive value (PPV) or (Precision):**

proportion of predicted positives which are actual positive[27].

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \ldots \ldots (3.4)
\]

• **F Score:**

Harmonic Mean of Precision and recall[27].

\[
\text{F Score} = \frac{2 (\text{Precision} \times \text{recall})}{\text{Precision} + \text{recall}} \quad \ldots \ldots (3.5)
\]
• **ROC analysis:**

Receiver Operating Characteristics (ROC) graphs are a useful and clear possibility for organizing classifiers and visualizing their quality (performance)[28]. A Roc curve is a plot of TPR vs FPR for different thresholds θ [29]. Receiver operating characteristic analysis is being used with greater frequency as an evaluation powerful methodology in machine learning and pattern recognition. The ROC is a well-known performance metric for evaluating and comparing algorithms.

• **Area under curve (AUC):**

AUC The area under the ROC is between 0 and 1 and increasingly being recognized as a better measure for evaluating algorithm performance than accuracy. A bigger AUC value implies a better ranking performance for a classifier [28].

3.3 Interfaces:

![Figure: 3.3 Import data](image)

The figure3.3 explain how to import data in the matlab program.
This figure 3.4 explain the data without Y. This data is become prepare to make dimensionality reduction using dimensionality reduction toolbox in matlab (drtoolbox).
The figure 3.5 explains the Locality preserving projection or principal component analysis command in drtoolbox. After applying the two algorithms PCA and LPP, the data is saved to apply cross validation code to evaluate the model.

Figure : 3.6 Cross validation

This figure 3.6 explains the cross validation code to evaluate the model using ten folds.
RESULTS AND DISCUSSION

4.1 Introduction:
This chapter explains the results of this research on five datasets using PCA and LPP dimensionality reduction algorithms, and then applied cross validation method to this new reduced features space using 10 fold to evaluate the model and then apply logistic regression to classifier the reduced data. All the performance measures: accuracy, sensitivity, specificity, f-score, precision is computed and the ROC analysis is plotted using spss software and then compute the area under the curve.

4.2 Results:

Table 4.1: The results of the performance measures for logistic regression with all variables

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Performance Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Climate Model Simulation Crashes (540 ×18)</td>
<td>0.9259</td>
</tr>
<tr>
<td>Heart (270 ×13)</td>
<td>0.7778</td>
</tr>
<tr>
<td>Spam base (4601×57)</td>
<td>0.9067</td>
</tr>
<tr>
<td>Phishing Websites (2456 ×30)</td>
<td>0.9283</td>
</tr>
<tr>
<td>Musk (Version 1) (476 ×186)</td>
<td>0.7358</td>
</tr>
</tbody>
</table>

The table 4.1 explain the result of the performance measures for logistic regression with all variables without using any algorithms.
Table 4.2: The results of the performance measures for logistic regression with principle component analysis algorithm

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Accuracy</th>
<th>Specify</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Model Simulation Crashes (540 × 18)</td>
<td>0.9444</td>
<td>0.5000</td>
<td>0.9800</td>
<td>0.9608</td>
<td>0.9703</td>
</tr>
<tr>
<td>Heart (270 × 13)</td>
<td>0.8148</td>
<td>0.8333</td>
<td>0.8200</td>
<td>0.8571</td>
<td>0.8276</td>
</tr>
<tr>
<td>Spam base (4601 × 57)</td>
<td>0.9197</td>
<td>0.9296</td>
<td>0.9040</td>
<td>0.8889</td>
<td>0.8964</td>
</tr>
<tr>
<td>Phishing Websites (2456 × 30)</td>
<td>0.9323</td>
<td>0.9329</td>
<td>0.9314</td>
<td>0.9048</td>
<td>0.9179</td>
</tr>
<tr>
<td>Musk (Version 1) (476 × 186)</td>
<td>0.8113</td>
<td>0.8438</td>
<td>0.7619</td>
<td>0.7619</td>
<td>0.7619</td>
</tr>
</tbody>
</table>

The table 4.2 explain the result of the performance measures for logistic regression with principle component analysis algorithm.
The table 4.3 explain the result of the performance measures for logistic regression with locality preserving projection algorithm.
4.3 Receiver Operating Characteristics (ROC) Curve Results:

4.3.1 Climate Model Simulation Crashes:

Figure 4.1: ROC Curve for Climate Model Simulation Crashes Dataset

The figure 4.1 explains the ROC curve for climate model simulation crashes data set.
4.3.2 Heart Data Set:

The figure 4.2 explains the ROC curve for the heart data set.
4.3.3 Spam base Data Set:

Figure 4.3: ROC Curve for Spam Base Dataset

The figure 4.3 explains the ROC curve for spam base data set.
4.3.4 Phishing Websites Data Set:

Figure 4.4: ROC Curve for Phishing Websites Dataset

The figure 4.4 explain the roc curve for phishing websites data set.
4.3.5 Musk (Version 1) Data Set:

![ROC Curve for Musk (version 1) Data Set]

Figure: 4.5  Roc curve for Musk (version 1) Dataset

The figure 4.5 explain the roc curve for musk (version1) data set.

4.5 The Area Under the Roc Curve (AUC):

<table>
<thead>
<tr>
<th>Datasets</th>
<th>PCA</th>
<th>LPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Model Simulation Crashes</td>
<td>0.912</td>
<td>0.948</td>
</tr>
<tr>
<td>Heart</td>
<td>0.812</td>
<td>0.904</td>
</tr>
<tr>
<td>Spam base</td>
<td>0.847</td>
<td>0.966</td>
</tr>
<tr>
<td>Phishing Websites</td>
<td>0.873</td>
<td>0.987</td>
</tr>
<tr>
<td>Musk (Version 1)</td>
<td>0.792</td>
<td>0.899</td>
</tr>
</tbody>
</table>

The table 4.4 explain the area under curve for PCA and LPP methods.
4.6 Discussions:

I used dimensionality reduction toolbox (drtoolbox) in MAT LAB software to select the features using PCA and LPP. All the computations are carried out on MATLAB (R2012a). The ROC curve and area under the curve are calculated by using the statistical package SPSS.16.0 (SPSS 14.0). The results show that using several performance measures with different data sets can help in understanding and comparing the performance of the algorithms. This research compared the performance of PCA and LPP performed on five sets of data on a regular basis. Table 4.1 the results of the performance measures for logistic regression with all variables, while Table 4.2 the results of the performance measures for logistic regression with principle component analysis. Table 4.3 the results of the performance measures for logistic regression with locality preserving projection. From those tables notice the locality preserving projection (LPP) method it has given a better result in all data sets although there are different in the number of Instances, number of attributes and type of attributes if compare to the principle component analysis (PCA) method and in all performance measures (accuracy, sensitivity, Specificity, precision, f-score and roc curve) LPP performs better than PCA. The ROC curves of PCA, LPP with all data set are shown in Figs.4.1-4.5 and LPP seems to be the best one. Table 4.4 represents the AUCs of each data set and the value of the area under the curve in LPP bigger than the value in the PCA and that indicates to LPP is better than PCA. Comparing to PCA method which it preserve the global structure, the LPP method preserving local structure which is more important than the global structure for many reason: it is important to maintain the intrinsic information of high-dimensional data when they are transformed to a low dimensional space for analysis, a single characterization, either global or local, may be insufficient to represent the underlying structures of real world data and the local geometric structure of data can be seen as a data dependent regularization of the transformation matrix, which helps to avoid over fitting, especially when training samples are scarce.
CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This research propose dimensionality reduction algorithm called LPP and then compare it with another method of dimensionality reduction approach called Principle component analysis for LR classification. The comparison includes several performance measures, which resulted in a valid and reliable conclusion. The performance of these approaches is evaluated in terms of accuracy, sensitivity, specificity, F-score, precision, AUC and ROC analysis. The comparison is done through experiments conducted on various types/sizes of data sets. The comparison shows that LPP it gives relatively good result in feature reduction and computational complexity when the training data size is relatively larger in comparison to the number of features. In LR, the features are required to be uncorrelated but not needed to be independent, when PCA and LPP are applied to the data sets with the number of features quite bigger than the data size, the dimension needs to be reduced to a very low dimension, large number of features lower reduced dimension, this resulted in loss of more information.

It can be stated that LR has proven to be a powerful classifier for high dimensional data sets and it also gives good efficiency when using the features selection methods, however, LPP seems to be the best method.

5.2 Recommendations

This research can be extended to other data mining techniques like clustering, association, It can also be extended for other classification algorithm such as neural network, decision tree and support vector machine and much more datasets should be taken. Moreover, This research recommends by using more than mathematical model to obtain the best results.
REFERENCES


[31] Yun Tang and Richard Rose, 2004, A STUDY OF USING LOCALITY PRESERVING PROJECTIONS FOR FEATURE EXTRACTION IN SPEECH RECOGNITION, collaboration with the DIVINES FP6 project and supported under NSERC Program Number 307188.


[35] Wei Li, Student Member, IEEE, Saurabh Prasad, Member, IEEE, James E. Fowler, Senior Member, IEEE, and Lori Mann Bruce, Senior Member, IEEE, 2012, Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 50, NO. 4.